A CUSTOMER SERVICE TEXT LABEL RECOGNITION METHOD BASED ON SENTENCE-LEVEL PRE-TRAINING TECHNOLOGY

Xiaoyu Qi[1,2], BoCheng[1,2], Kang Yang[3], Lili Zhong[3] and Yan Tang[3]

[1]Shenzhen Audencia Business School, We Bank Institute of Fintech, Shenzhen University
[2]State Key Laboratory of Network and Switching Technology of Beijing University of Posts and Telecommunications
[3]Ping An Bank Co. Ltd.

ABSTRACT

Customer service text data is known as the dialogue text data between users and customer service provider, and it contains a large amount of user information. The effective use of customer service text content can bring great business plan optimization to the service provider. Based on the traditional machine reading comprehension model, this paper builds a customer service text user's attribute label recognition model, and proposes a model pre-training method based on sentence-level pre-training technology: aiming at the background of poor performance of the model in answering comprehensive full-text content analysis questions such as user intent and text sentiment analysis. This paper extracts text summaries based on the T5-pegasus model, constructing a text summaries dataset for model pre-training. Then build a text summarization model including an ERNIE pre-training model, train the model's ability to understand the full text, and improve the model's ability to answer questions that need to be combined with full-text content understanding, such as user intent and sentiment analysis. Use the pre-trained model to solve customer service text label recognition tasks based on machine reading comprehension tasks. The test results based on the data set show that the improved model has an improvement in performance of customer service text label recognition task.

KEYWORDS

Machine Reading Comprehension, Pre-trained model, Customer Service Text Analysis, Natural Language Generation, Attention Mechanism

1. INTRODUCTION

The problem of customer service text label recognition mainly describes the construction of a model to identify the user attribute content contained in the customer service text, given the user attribute labels contained in the customer service text. This kind of problem can be modelled as selective reading comprehension in machine reading comprehension. As one of the important research directions of reading comprehension, selective reading comprehension has been widely used in real life, such as text analysis, intelligent question answering, and customer service text application to optimize targeted recommendation content for customers. There are various problems in customer service text label extraction problems, and sometimes there is a problem of

David C. Wyld et al. (Eds): SIPM, ITCA, CMIT, FCST, CoNeCo, SAIM, ICITE, ACSIT, SNLP -2023
pp. 139-146, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130915

grasping and analysing the overall intention of the customer or the customer's dialogue attitude. However, traditional reading comprehension models often only consider partial text semantics when answering questions and cannot grasp the full text well. To solve this problem, this paper proposes a model optimization strategy based on sentence-level pre-training technology, which can train the full text of the model. Comprehension ability, and then improve the performance of the model. The main research contents of this paper are as follows: 1) Aiming at the problem of poor understanding of the full-text content of the model, it is proposed that the full-text grasping ability of the model can be trained; 2) Based on the T5-pegasus model, extract the summary content of the customer service text, and build a training model for the ability to read the full-text Customer service text summary data set; 3) Based on the customer service text summary data set, build a summary generation model with the ERNIE model as the pre-training model, and train the ERNIE model to understand the full-text content; 4) The test results of different data sets show that the proposed The sentence-level pre-training technology can effectively improve the model's ability to extract customer service text user attributes.

2. RELATED WORK

The machine reading comprehension task is based on human beings' actual reading text content understanding and answering related questions. In this task scenario, the machine will accept the input article content, question content, and option content to answer questions and give answer options. Selective reading comprehension can choose a most suitable answer from the alternative options based on the given article content, question content, and option content. At the same time, the selective reading comprehension task includes a lot of reasoning and inductive tasks combined with external common sense, which can more comprehensively grasp the full text content. Dataset for selective reading comprehension has a CLOTH **Error! Reference source not found.**, OpenBookQA**Error! Reference source not found.**, RACE**Error! Reference source not found.** and C3**Error! Reference source not found.**, etc. In addition to providing articles and questions, OpenBookQA also provides an external knowledge base for answer selection. When answering each question, the answer is selected by combining the facts of the external knowledge base. Therefore, when using this data set, in addition to dealing with data Integrating, it is also necessary to accurately utilize the external knowledge input into the model. RACE is currently one of the most widely used selective reading comprehension datasets. It consists of high school English reading comprehension questions. The RACE dataset contains many complex questions, involving text content understanding and reasoning. There are roughly five types of reasoning questions in it, which are detailed reasoning, article summary, global reasoning, basic common sense and attitude analysis, so it can effectively evaluate whether the reading comprehension model has text reasoning ability, and can evaluate the ability of the reading comprehension model. The C3 data set is a Chinese reading comprehension data set. The articles come from the content of the Chinese Proficiency Test and the National Chinese Test. It collects multiple-choice questions in a more flexible form. The data is divided into formal written text and oral text. The data samples are short text, so they perform poorly when training on long text tasks.

Based on the above data sets, there are more reading comprehension models based on neural network structure. Currently, the most widely used pre-training model in the industry is the BERT model **Error! Reference source not found.**. The BERT model was open sourced in 2018. It uses a bidirectional transformer**Error! Reference source not found.**and an attention mechanism to train a language understanding model, and uses a random character mask to understand the semantics of text fragments; Zhang**Error! Reference source not found.** et al. proposed a method for multiple-choice reading comprehension, the dual co-matching network (DcMN+) model. The magnitude BERT model **Error! Reference source not found.**

ALBERT model, the XlNet**Error! Reference source not found.**model proposed by Yang et al., and the RoBERTa**Error! Reference source not found.** model proposed by Liu et al. have good performance for different downstream tasks. The ERNIE model **Error! Reference source not found.** is Baidu's The open-source high-performance text understanding model in 2019, adding masking technology for phrases and entities on the basis of BERT, and adding continuous multi-task learning technology, which has a strong comprehensive ability in text analysis task solving.

Sentence generation technology also relies on deep learning network construction models. Currently, the most commonly used deep pre-training models in the field of generative text summarization include MASS**Error! Reference source not found.**, TAAS**Error! Reference source not found.**, UniLM**Error! Reference source not found.**, T5**Error! Reference source not found.**, STEP**Error! Reference source not found.**, BART**Error! Reference source not found.**, PEGASUS**Error! Reference source not found.**, ProphetNet**Error! Reference source not found.**' etc. These technologies promoted the development of text generation from the perspectives of modelling methods, model size optimization, and corpus optimization.

3. TEXT SUMMARY DATASET EXTRACTION TECHNOLOGY BASED ON T5 MODEL AND GSG TECHNOLOGY

3.1.T5 Model Technology Exploration

In the field of generating text summarization, the T5 model has excellent performance. The T5 (Text-to-Text Transfer Transformer) model structure is still an Encoder-Decoder structure stacked by Transformer layers. The Decoder structure is very similar to the Encoder structure, but the Decoder has a standard attention layer after the self-attention layer. This standard attention layer will take the output of the Encoder into the attention calculation.

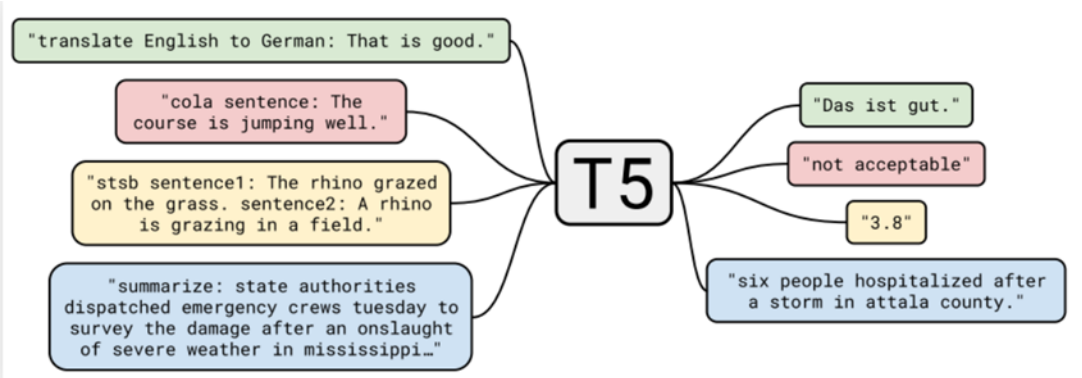

Figure 1. Illustration of T5 model's missions.

T5 is very similar to the original Transformer structure, the difference lies in: the author uses a simplified version of Layer Normalization, removes the bias of Layer Norm; puts Layer Norm outside the residual connection; position coding: T5 uses a simplified version of the relative position encoding, each positional encoding is a scalar, is added to the logits for computing the attention weights. The positional encodings are shared across layers, but within the same layer, the positional encodings of different attention heads are learned independently. A certain number of position Embeddings, each corresponding to a possible key-query position difference.

For unified modelling, all tasks are converted to a "text-to-text" format, and the model is optimized by maximizing the likelihood objective (teacher-forcing), keeping the goals of pre-training and fine-tuning unified. To let the model know what task it is performing, prefix the input text with a task-related prefix. The authors found that changing the wording of the prefixes had little effect on the results, so the effect of different prefixes was not explored.

The T5 model architecture adopts the Encoder-Decoder architecture, and the mask method adopts part without mask and part with diagonal mask. The difference between different models is mainly in the mask design in Transformer's self-attention. The mask design strategy determines which contextual information will be involved in the calculation of self-attention. The following figure shows the three most common types of masking strategies: First, no masking, the context information on the left and right sides are involved in the calculation of attention, suitable for bidirectional language models such as BERT in the middle: conventional diagonal masking Code, only the previous token will be used as context information to participate in the attention calculation, suitable for one-way language models such as GPT. Some without masks, some with diagonal masks, suitable for seq2seq languages such as encoder-decoder model, the output sequence in the Encoder and the previous tokens in the Decoder will participate in the attention calculation of the following tokens.

3.2. GSG Technology

GSG technology is a word masking method different from MLM. It is assumed that the closer the pre-training target is to the downstream task, the faster and better the fine-tuning effect will be. Therefore, more inter-semantic relationships can be learned and by randomly masking out some sentences, the relationship between sentences can be understood and new text content can be generated through the relationship between sentences. To this end, the downstream language task is described as "extracting text summarization", while the pre-training goal is to generate interstitial sentences. Afterwards, a pre-training model is proposed to generate text summaries. The key point is that the goal of pre-training is to generate gap sentences, so when extracting text summaries, a simple fine-tuning model has a great performance improvement. In the actual experiment process, some single sentences in the original text are directly masked, and then the remaining gap sentences are directly spliced as a text pseudo-summary. The masked sentences are replaced by the [Mask] symbol. In order to get closer to the fine-tuning of downstream tasks, we will select more important sentences in the article for masking. GSR (gap sentence ratio) is a hyper parameter of GSG (Gap Sentences Generation), which refers to the number of selected interval sentences in the document divided by the total number of sentences, which is equivalent to the mask ratio in other studies.

Use 3 strategies to select gap sentence
1.Randam, uniformly randomly select m sentences.
2.Lead, select the first m sentences.
3.Principal, and select top-m most important sentences according to importance.

Sentence A and the RANGE-F1 of the remaining documents except the sentence A are used as the calculation index of importance. The formula is as follows:

$$s_i = rouge(x_i, D \setminus \{x_i\}), \forall i$$

The way of taking sentences according to the importance is considered from the following two aspects:

Table 1. Ways of calculating ROUGE

Ways of taking sentences	Ways of calculating ROUGE
Ind	Uniq
Seq	Orig

Independent (Ind): When comparing rouge, each time one sentence is compared with the remaining sentences to get the score, and finally the score of each sentence is obtained, and the sentences with the highest scores are obtained.

Sequence (Seq): If you want to take the most important 2 sentences from the 4 sentences this time, take (12, 13, 14, 23, 24, 34) sentences, compare with the remaining sentences to get the score, and finally get the group with the highest score (such as 12 example)

Unique (Uniq): When computing the amount of n-gram in rouge1 score, consider all n-gram as nonrepeatable set. For example, there are three n-grams in "我爱哈哈哈哈".
Origin (Ori): The original method of computing amount of n-gram. There are six n-grams in "我爱哈哈哈哈".

In this formula, each sentence is calculated independently and the top-m, called as Ind, is selected from it. The researchers also believe that by maximizing the ROUGE1-F1 between the selected sentence set and the remaining documents based on greedy thinking, the top-m can be selected in an orderly manner, that is, Seq. The specific algorithm is as follows, where S is the selected A collection of sentences, D is the collection of all sentences in the document.

Table 2. Pseudocode of GSG Algorithm

Algorithm 1 Sequential Sentence Selection

$S := \emptyset$
for $j \leftarrow 1$ *to* m **do**
$s_i := rouge(S \cup \{x_i\}, D \ (S \cup \{x_i\})$
$\forall i \ s.t. x_i$ *not belong to* S
$k := argmax_i\{s_i\}_n$
$S := S \cup \{x_k\}$
end for

When calculating ROUGE-F1's value, consider n-grams as a collection, i.e., Uniq. Or calculating repeatedly the same n-gram as the original set, i.e., Orig. There are four ways of choosing the Principal, the optional parameters are Ind/Seq and Orig/Uniq.

Generating customer service text summaries based on the T5 model, the loaded T5 model is T5-pegasus. This model adds the word segmentation function to BERT's tokenizer, improves the word segmentation table, and adds the first 100,000 words after stuttering word segmentation to the original Chinese BERT character token dictionary, and then modifies the internal structure logic of the tokenizer so that the tokenizer can segment the vocabulary. Then use the modified tokenizer to segment the pre-training corpus for training, and count the frequency of each segmented word, and finally only keep the 50,000 words with the highest frequency, and get a final vocabulary size of 50,000 Word segmentation table to build the final tokenizer used. At the same time, it draws on the ideas of the PEGASUS paper. The idea is to extend the mask level to sentences, that is, for an article, some sentences are masked out through some strategies, and then the remaining sentences are used to predict the content of the masked sentences. T5-PEGASUS

draws on similar ideas, the difference is that it designs a new pre-training sample construction method. For a document, through a search strategy, a certain number of target sentences are selected, and use the remaining sentences as source sentences to do the seq2seq task.

Then load the chapter content of the customer service text dataset, and pass it into the T5 model as an input to generate a summary of the chapter content, and edit and construct the summary dataset according to the article number. Train the ERNIE model based on the text summary data set, and build a text summary model based on the ERNIE pre-training model. The model training process uses the text content as the task input and the summary content as the output. In the T5-PEGASUS paper, GSG is used to generate gap sentences. It mainly considers sentences can be extracted as a pseudo-summary, some important sentences are selected instead of MLM tasks similar to the BERT model, and some word masks are randomly selected.

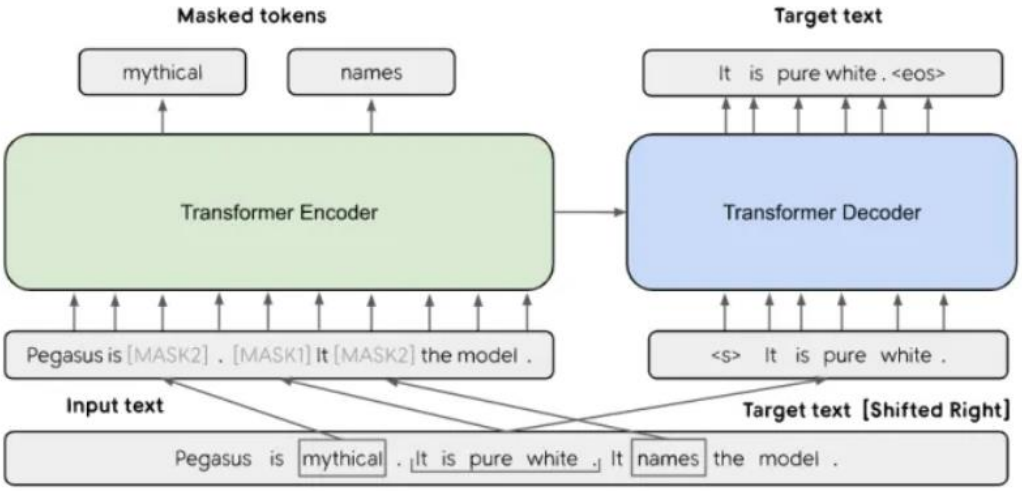

Figure 2. Illustration of Pegasus Model

4. MODEL PRE-TRAINING AND EXPERIMENTAL RESULTS EVALUATION BASED ON SUMMARY DATASET

First build the T5-PEAGSUS model, which uses masking entire sentences from the document and generating these empty sentences from the rest of the document can be a good pre-training target for downstream summarization tasks. In particular, selecting sentences that are putatively important outperforms bootstrapped or randomly selected sentences. We hypothesize that this objective is applicable to abstract summarization, as it is very similar to downstream tasks, encouraging understanding of whole documents and generation of similar summaries. We refer to this as Self-Supervised Target Gap Sentence Generation (GSG).

The experiment uses the ERNIE model for ablation experiments. The basic ERNIE model configuration is the same, and ERNIEBASE-3.0 is used as the pre-training model. The model contains 12 layers of computing units, 768 hidden units, 12 heads, and 118M parameters. The ERNIEA model adds text summary generation technology on the basis of the ERNIE model, so that the encoder of the model has the ability to overview the full text and improve the accuracy of answering user intent questions. The T5 model generated 2998 text summaries as a text summaries dataset for subsequent training.

The ERNIE pre-training model takes the text summary and the original content as input, uses the ERNIE model to obtain the original text correlation parameters, and then converts the text summary results generated by the output model through the text generation module, compares it with the target summary text to calculate the loss, and feeds back to update the model parameters. The ERNIE model outputs the parameters used to generate text summaries in the pre-training task, and the parameters are passed into the text generation module. The text generation module receives the pre-training model parameters and generates new text according to the previously generated text content and parameters, so that the ERNIE model can obtain the ability to control the full text.

The results of the experiment are shown in the table. It can be found that the ERNIEA and ERNIEA models pre-trained by the summary generation task have improved in the C3 data set and the customer service text user attribute label extraction task.

Table 3. Results of the Tag Recognition Experiment (%)

	C3	Customer Service Text Data Set
Albert-base	59.6	53.5
BERT	64.5	57.3
Roberta	67.5	61.6
ERNIE	73.7	67.2
ERNIEA	74.8	68.3

5. CONCLUSIONS

Based on the customer service text user label recognition data set, this thesis uses the ERNIE model to construct a customer service text user attribute label content recognition model on the basis of investigating machine reading comprehension models. Considering that the model is limited to local text semantics when answering user intentions, sentiment analysis, etc. There is no overview of the full text, and a sentence-level model pre-training technology is proposed. Using the generated text summary data set to train the model has the ability to grasp the full-text content. At the same time, comparison and ablation experiments are carried out to prove the effectiveness of the sentence-level pre-training technology.

ACKNOWLEDGEMENTS

This work is supported by Swift Fund Fintech Funding.

REFERENCES

[1] Xie Q, Lai G, Dai Z, et al. Large-scale Cloze Test Dataset Created by Teachers[C]. EMNLP 2018.
[2] Mihaylov T, Clark P, Khot T, et al. Can a suit of armer conduct electricity? A new dataset for open book question answering[J]. 2018.
[3] Talmor A, Herzig J, Lourie N, et al. CommonSenseQA: A question answering challenge targeting commensense knowledge[C]. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, Volume 1, Pages 4149-4158, 2019
[4] Sun K., Yu D., Yu D., et al. Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension[J]. arXiv:1904.09679.
[5] Devlin J, Chang M. W., Lee K., et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, Volume 1, Pages 4171-4186, 2019

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [OL]. arXiv:1706.03762. https://arxiv.org/pdf/1706.03762

[7] Zhang S.，Zhao H.，Wu Y., et a1．DCMN+：dual co—matching network for multi-choice reading comprehension[C]．Proceedings of the AAAI Conference on Artificial Intelligence，2020, 34(5):9563—9570．

[8] Lan Z., Chen M., Goodman S., et al. A lite BERT for self-supervised learning of language representations. [OL] https://arxiv.org/pdf/1909.11942.pdf.

[9] Yang Z., Dai Z., Yang Y., et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding [OL]. https://arxiv.org/pdf/1906.08237.pdf

[10] Liu Y., Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach [OL]. https://arxiv.org/pdf/1907.11692.pdf

[11] Sun Y., Wang S., Li, Y., et al. ERNIE: Enhanced Representation through Knowledge Integration. [C]. arXiv preprint arXiv:1904.09223, 2019.

[12] Song K., Tan X., Qin T., et al. MASS: masked sequence to sequence pre-training for language generation. [OL]. (2019-05-13) [2021-02-10]. http://arxiv.org/pdf/1905.02450v3.pdf.

[13] Zheng C., Zhang K., Wang H. J., et al. Topic-aware abstractive text summarization. [OL]. (2020-10-20) [2021-02-10]. https://arxiv.org/pdf/2010.10323.pdf.

[14] Li Dong, Y ang Nan, Wang Wenhui, et al. Unified language model pre-training for natural language understanding and generation. [OL]. https://arxiv.org/pdf/1905.03197.pdf.

[15] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[EB/OL].https://arxiv.org/pdf/1910.10683.pdf

[16] Zou Y., Zhang X., Lu Wei, et al. Pre-training for abstractive document summarization by reinstating source text[C]. //Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. [S. 1.]: Association for Computational Linguistics, 2020:1-5.

[17] Lewis M, Liu Y., Goyal N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meetings of the Association for Computational Linguistics. [S.1.]: Association for Computational Linguistics, 2020:1-5.

[18] Zhang J., Zhao Y., Saleh M., et al. PEGASUS: Pre-Training with extracted gap-sentences for abstractive summarization. [C] 37th International Conference on Machine Learning, ICML 2020, Volume PartF168147-15, Pages 11265-11276, 2020.

[19] Qi W., Yan Y., Gong Y., et al. ProphetNet: predicting future n-gram for sequence-to-sequence pre-training[C]//Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing.[S.1.]: Association for Computational Linguistics, 2020:1-5.